2014 DLF Forum

Managing the Digitization of Large Press Archives
Monday, October 27, 3:45am-5:15pm
Salons 4,5,6, Georgia Tech Hotel and Conference Center

Colocated with: **Audio and Video at Scale: Indiana University's Media Digitization and Preservation Initiative** (Group notes: https://docs.google.com/document/d/1Y9VK9HuDbZa6fs6gGx7teeEb4vsTjgeURTQ3clfvrqY/edit#)

Colocated with: **Building a Ten-Campus Digital Library Collection at the University of California** (Group notes: https://docs.google.com/document/d/1tOcC09wNylyaRDXhd68egnSLp5Nrwqq5-I8oVDRudVk/edit#)

Session Leaders
Bassem Elsayed, Bibliotheca Alexandrina
Ahmed Samir, Bibliotheca Alexandrina

Slides
http://www.slideshare.net/DLFCLIR/managing-the-digitization-of-large-press-archives-small

Notes

Review of digitization of CEDEJ press article collection; multiple modules to index, control data entry, digitization, review of articles,, Collection includes over 800,000 press clips, Arabic, English and French; so far, over 100,000 articles online
Organized by folder, file, sub-file hierarchy; multi-page article organization; 6-phase workflow: Physical data preservation, (2) Digitization (digital assets factory) (3) Processing (4) Web site (5) Indexing data with Solr search engine (6) Cataloging
Digital assets factory: Manages digitization process; multiple phases: scanning, processing, OCR, and back-up archive,
Cataloging workflow: Define article, add or dupe or reject images in article, ; can manage article “creation” in graphical user interface using drag-and-drop workflow
System workflow: Website - articles, publishers, timeline
Article viewer: JSONP
Q: OCR for Arabic script - developed in Egypt by a software house there; very reliable software for Arabic script